DOCUMENT RESUME

ED 193 307 TM 800 622

AUTHOR Johnson, Eugene G.: And Others
TITLE Measures of Scorer Agreement and Patterns of Disagreement.
PUB DATE Mar 78
NOTE 30p.: Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests: *Educational Assessment: Elementary Secondary Education: Experimenter Characteristics: *National Surveys: Sampling: *Scoring: Situational Tests: Testing Problems: *Test Reliability
IDENTIFIERS *Interrater Reliability: *National Assessment of Educational Progress

ABSTRACT

The scoring scheme of the National Assessment of Educational Progress (NAEP) is studied, as well as techniques for estimating its reliability. NAEP data are collected on students of age 13 in the fall, age 9 in the winter, and age 17 in the spring. The scoring task, involving hand scoring of open-ended items, may vary between 25,000 responses to 10 items and 1,000,000 responses to 400 items. The size of the scoring staff may vary between two and 20 scorers, and their work is subject to 10 percent quality control over the entire scoring period. The general scoring scheme requires that responses be divided into two overall evaluative categories: acceptable and unacceptable. Acceptable responses provide some evidence that the respondent meets the relevant educational objective: unacceptable responses are not adequate to provide evidence. A discussion of techniques for estimating reliability of the scoring scheme involves: (1) determining the percent of observed agreement among scorers and comparing it to the agreement one would expect on the basis of chance alone: and (2) examining the patterns of disagreement in the scoring of individuals. The test statistics for measuring disagreement are applied to data from multi-way tables. (PL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED193307

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

MEASURES OF SCORER AGREEMENT AND PATTERNS OF DISAGREEMENT

by:
Eugene G. Johnson
Nancy W. Burton
Michael J. Noe

National Assessment of Educational Progress
March 1978

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

T. Pratt

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 800622

Presented at the 1978 Annual Meeting of the American Educational Research Association (Toronto) in a Round Table Discussion

Introduction

National Assessment of Educational Progress (NAEP) is a census-like assessment project which collects data on a national probability sample. NAEP has collected data in ten different learning areas from three in-school age classes (9-year-olds, 13-year-olds, 17-year-olds). The major purpose of NAEP is to measure changes across time in performance on objectives-referenced exercises. Many of the exercises NAEP uses are open-ended and must be hand scored. NAEP's hand scoring does not generally consist of assigning responses to points on an ordinal scale. Instead, responses are almost always assigned to nominal (descriptive) categories.

Because NAEP data are basically useful for comparisons -- either to determine the progress of education over time, which is NAEP's major goal or for state or local education agencies to compare local results with a national benchmark -- it is important that hand scoring not depend heavily on the scorer or the time of the scoring. It is known that when essays are scored for quality on ordinal scales, the scores vary with the context in which the papers are read (Coffman, 1971). If these findings are true also for NAEP's nominal data, measurement of change would require that all responses from all points in time be read in the same context and time. Each state or local replication would also entail the rescoring of NAEP national data. This is not only expensive; it is also likely that the NAEP benchmark performance level would change in each new scoring context which would be detrimental to NAEP's credibility for benchmark data. The purpose of this study, then, was to begin to answer the question: How reliable are National Assessment scoring procedures?

OCT 6 1980

Some Distinctions About "Reliability"

Questions about reliability of data can be asked at a variety of levels. At the most general level, for the assessment, one may ask how reliable are the results reported for the nation on any particular assessment item or group of items. How different would the results be if we had chosen a different sample or administered on different days or used a different set of scorers? Because any one national assessment result is based on the performance of at least 2,500 students from 150 schools in 150 geographical locations on any one of 40 days by 100 administrators and scored by anywhere from 2 to 20 scorers, the estimate of random variability -- the standard error[1] -- of the result gives an excellent answer to that general question.

The particular questions addressed by the present study are more specific than this general reliability question. Essentially, we are asking: Given the estimate of overall random variability obtained from the standard error calculation, what component of that variation is attributable to scoring? Even at this more specific level, there are several reliability questions that are worth asking.

First, there is the evaluative question: Is this measure precise enough to be used again, for an assessment of change?[2] Equally, a state or local assessment staff may ask whether a certain NAEP item is precise enough for its purposes.

---

[1] See Folsum (1977) for a discussion of how NAEP standard errors are computed.

[2] We assume that a prior question -- is this measure precise enough to use at all? -- has already been answered. Note, however, that the method for answering the present question will also answer the prior question.

Second, there are operational questions. For example, is there a scorer or group of scorers who differ from their colleagues? Are there certain categories of response that cannot be clearly distinguished? Are there certain item types or scorer types or training procedures or working conditions that affect precision?

The data for this study were gathered for operational reasons -- to improve overall scoring procedures, to detect problem scorers and either retrain or dismiss them; to detect problem items so the scoring guides could be revised or replaced. Answering the overall evaluative questions was only a secondary concern.

The Scoring Scheme Studied

The Working Conditions

National Assessment data are collected throughout an entire school year (13-year-olds in the fall, 9-year-olds in the winter, 17-year-olds in the spring). The hand scoring of open-ended items is done in parallel with the data collection so responses to any one item are being scored continuously over three months; if the item is administered at more than one age, the scoring may continue over nine months.

The scoring task may vary between 25,000 responses to ten items (there are 2,500 responses to each item) and 1,000,000 responses to 400 items depending on the learning areas assessed. The size of scoring staff has varied between 2 and 20 scorers. The scorers generally have some teaching experience in the area being scored. They are trained extensively before each age-level of work and their work is subject to ten percent quality control over the entire scoring period.

The Items

National Assessment uses a wide variety of open-ended and performance exercises. Students have been asked to write essays, play instruments, and to tell an interviewer about their reading habits. Each of these performances requires a unique scoring protocol and, consequently, might require a unique estimate of reliability. These unique exercises are not the subject of this study.

National Assessment staff has developed a general scoring form that fits the range of open-ended items most common in the assessment. For example, the exercises in the present study that fit under the general NAEP scoring scheme ranged from addressing an envelope to an essay defending one's position on the statement, "A woman's place is in the home."

The general scoring scheme for such exercises requires, at least, that responses be divided into two overall evaluative categories: acceptable and unacceptable. Acceptable responses provide some evidence that the respondent meets the relevant educational objective; unacceptable responses are not adequate to provide evidence.[3] However, the general scoring scheme does not stop with this simple 0,1 data. Within the set of "acceptable" responses, several descriptive categories of response are defined. For the "women's place" exercise, for example, the respondent's choice of supporting evidence was categorized. Did the respondent appeal to conventional wisdom, personal experience, a particular authority, literature or history, or to legal precedent? In mathematics computation exercises, unacceptable

[3]Even attitude and self-report items are classifiable as "acceptable" or "unacceptable" under this definition. For example, a political position would be classifiable under an objective such as: "The respondent can clearly state and logically defend his reasons for supporting a political candidate."

responses are often categorized as to the type of error made. (See Appendix A for several example scoring guides.)

Discussion of Techniques for Estimating Reliability

Reliability estimates can be seen as a special subset of all measures of association. In general, a measure of association tells whether one can predict B's behavior, given knowledge of A's behavior. A measure of reliability tells whether B's behavior and A's behavior agree. Light (1973) illustrates the difference between measures of association and measures of agreement with the following example.

Figure 1. Light's Illustration of the Difference Between Agreement and Association

a.

Teacher 1 \ Teacher 2	A	B	C	
A	25	10	15	50
B	10	4	6	20
C	15	6	9	30
	50	20	30	100

$\chi^2 = 0$

b.

Teacher 1 \ Teacher 2	A	B	C	
A	25	5	20	50
B	15	4	1	20
C	10	11	9	30
	50	20	30	100

$\chi^2 > 0$

The diagonal of the table shows the cells in which both teachers agree. Note that the diagonal elements are the same in both tables, but the χ^2 test of independence (Kendall and Stuart, 1967) is different for the two tables. Light (1973, p. 330) points out that:

> The reason for the nonzero chi-square in b is not that the observed agreement differs from the expected agreement under H_0; rather it is because the pattern of disagreements is different from expectation. The difference in the disagreeing cells should not affect a measure of agreement, and chi-square is therefore inappropriate for this problem.

In general, the differences among reliability estimates could be characterized by the amount of emphasis placed on these diagonal elements. It should be noted that the purpose of measures of agreement is to answer the evaluation question, as opposed to the operational or quality-control question. For many quality-control applications, including the ones we discuss in the p esent study, the emphasis is almost entirely on the off-diagonals or on disagreements. Even when evaluating agreements, however, some attention must be paid to off-diagonal elements for the case of ordered scales. For example, given an ordered scale from one to ten, responses placed in the cell eight, nine are in greater agreement than responses placed in the seven, ten cell. When evaluating agreement among nominal categories -- as in the Light example -- only the on-diagonal elements should be considered.

Even though we are primarily concerned with the operational or quality-control aspects of reliability in this apper, quality-control is a two-step process. First, one decides whether the overall level of agreement is satisfactory, then one analyzes the unsatisfactory data to locate the particular problem. In the next section we discuss various methods of determining the overall level of agreement.

Measures of Agreement

The simplest method of determining agreement among scorers is simply to determine the percent of agreement -- the proportion that the diagonal elements are of the whole:

$$(1) \quad \frac{1}{n} \Sigma n_{ii}.$$

Many more complex formulas have been proposed. They basically share the same purpose: to compare this observed agreement to the agreement one

would expect on the basis of chance alone.

There are several factors one must consider in selecting one of these more complex measures of agreement. First, can we assume that there is a continuous variable underlying the categories or not? If there is, can we assume anything about the distribution of that variable? If there is not, can we at least assume the categories to be ordered, or must we decide to treat the categories as completely nominal?

Second, can we assume that the set of scorers being compared are drawn from the same population, or must we allow for non-random difference among scorers? Must we allow for a biased preference or avoidance of any category by any scorer?

Finally, we must consider the distribution of true performances in the population to be measured. Is there any reason to believe that the true marginal distribution of performances will be seriously disproportional; namely, nearly all performances belonging to one category?

Answering these three sets of questions helps us select a correct measure of agreement.

Question 1: What kind of order exists among the categories? There is some order inherent in the NAEP scoring scheme since all responses are initially divided into acceptable/unacceptable strata. However, there is clearly no meaningful underlying metric, so reliability statistics based on Pearson product-moment correlations or upon analysis of variance techniques[4] can be ruled out. Since the NAEP scoring scheme includes unordered

[4]"In those cases in which all the polytomies of a study arise jointly from a multivariate normal distribution or an underlying continuum. one would naturally turn to measures of association based on the correlation coefficient." (Goodman and Kruskal, 1954, p. 736)

categories within the ordered acceptable/unacceptable strata, the purpose of the computation will determine whether the categories are treated as ordered 0, 1 data or as unordered polytomies.

When NAEP items are considered for the purpose of summaries (the average proportion of success on items dealing with political knowledge, for example) or for reporting educational trends (in overall science achievement, for example), exercises are used as 0, 1 ordered scales. It is also useful to treat exercises as 0, 1 scales when evaluating the generalizability of the results over such operational variables as time of scoring, time of administration, position in package (item booklet), item format, and so on. That is because convenient and powerful analysis of variance techniques exist for evaluating generalizability (Cronbach, Gleser and Rajaratnum, 1972).[5] The appropriateness of using analysis of variance techniques for 0, 1 data was addressed by Cochran (1950, p. 262) as follows:

> If the data had been measured variables that appeared normally distributed, instead of a collection of 1's and 0's, the F-test would be almost automatically applied as the appropriate method. Without having looked into the matter, I had once or twice suggested to research workers that the F-test might serve as an approximation even when the table consists of 1's and 0's. As a testimony to the modern teaching of statistics, this suggestion was received with incredulity, the objection being made that the F-test requires normality, and that a mixture of 1's and 0's could not by any stretch of the imagination be regarded as normally distributed. The same workers raised no objection to a χ^2 test, not having realized that both tests require to some extent an assumption of normality, and that it is not obvious whether F or χ^2 is more sensitive to the assumption.

See also D'Agostino (1971) for a more complete discussion of the topic.

---

[5]See, for example, Burton and Johnson (in preparation) for a discussion of generalizability over scoring conditions and Burton (1976) for a discussion of generalizability over item format and position for NAEP.

For the present purpose of scoring quality control, it makes sense to evaluate the entire set of categories, not only to determine the reporting potential for each category, but also to determine as precisely as possible the particular source of disagreement among scorers. Taken as an entire set, the scoring scheme must be considered an unordered or nominal scale. This decision eliminates those reliability estimators that assume order, though not necessarily an underlying continuum, such as Spearman's Rho (Glass and Stanley, 1970).

Question 2: Are the scorers unbiased? Krippendorff (1970) distinguished among reliability coefficients on two dimensions. The first is whether the data are ordered or unordered (question 1, above); the second is on the extent of agrement expected among scorers. In the most stringent case, one expects both scorers' work to estimate the true marginal distribution of responses. Thus the "expected" or chance-level call frequency can be computed on the joint marginal distribution, that is, on the average frequencies in categories i and j:

$$(2) \quad e_{ij} = \frac{1}{n_{++}} \left(\frac{n_{i+} + n_{+i}}{2}\right) \left(\frac{n_{j+} + n_{+j}}{2}\right)$$

where

e_{ij} = expected number in the i, j cell

n_{i+} = the marginal frequency of category i summed over scorer 1 (columns)

n_{+j} = the marginal frequency of category j summed over scorer 2 (rows)

n_{++} = total number of responses examined by both scorers

If we allow the marginal distribution to depend on the scorer -- if we allow for the possibility that a scorer may have some unique category preferences -- then the expected cell frequency is expressed:

$$(3) \quad e_{ij} = \frac{1}{n_{++}} (n_{i+}) (n_{+j})$$

Clearly, scorers may have a unique preference for some category. NAEP staff has found it very difficult to train scorers to abandon such preferences -- our attempts have generally led to overcompensation. Therefore, we have developed scoring procedures that randomize responses over scorers to control the effects of mild scorer preferences.

This second decision, to tolerate mild scorer preferences, eliminates coefficients like Scott's (1955) coefficient of inter-coder agreement or the intraclass correlation coefficient (Pearson, 1901), which assume a joint marginal distribution. (See Krippendorff, 1970, for a full discussion of this distinction.)

In summary, then, NAEP requires a measure of agreement for quality-control purposes which does not require ordered scoring categories and does not require complete marginal agreement among scorers. So far as we can determine, there are basically two measures of agreement that fit these crtieria: the most simple measure -- the percent of responses on which both scorers agree -- and the measure called kappa, proposed by Cohen (1960). The choice between these two measures depends on one's answer to the next question.

Question 3: Are the responses distributed reasonably proportionally among scoring categories? In classical tests, there is no point in including a question for which all responses fall into one category since such items do not discriminate among respondents. In assessment and survey work, such items do appear. It is of great interest, for example, to know that nearly all voters favor some policy or that almost no students

make spelling errors in a short essay. The information is of particular interest if it is tracked over long periods of time. Even though the results shows no variation in one assessment, they may vary over time.

None of the more complex coefficients of agreement are appropriate when there is little variation in the data, since the chance level of agreement begins to approach the maximum possible agreement. One may try some data transformation such as rescaling the marginal categories so they are equiprobable (Goodman and Kruskal, 1954, p. 746) to overcome the disproportionality, but a more basic problem remains. This problem is well defined by Davidson (1973, p. 1):

> Coding tasks sometimes have as their primary objective the detection of a relatively rare event. In such cases, the coder may be presented with tens of open-ended responses before encountering a single instance of this rare event. There is a danger here that after so many presentations of the "non-event" the coder may become insensitive to the event when it really is presented.

In this case, one is not so concerned about disagreements among scorers as artificial agreement due to inattention. No data adjustment procedure or complex estimator will detect these false agreements. For these rare-case items, we feel that the simple percent of agreement statistic is quite an accurate measure of agreement. We attempt to solve the problem of false agreement managerially: by providing short (six hour) work days, by providing a variety of scoring tasks, and by letting the scorers know their work will be checked. For the vast majority of NAEP items -- those for which no more than eighty percent of all responses fall into one category -- Cohen's kappa provides an excellent initial estimate of agreement.

Examination of the level of scorer agreement -- estimated either by Cohen's kappa or the simple percent of agreement -- will identify a certain number of items for which the level of scorer agreement is

unsatisfactory. The next problem is to identify the specific problem with the item. We would like to ascertain whether the disagreements are occurring in a random manner or if these disagreements have a definite pattern which can be ascribed to some problem in the scoring. For example, we would like to detect a scorer (or group of scorers) who frequently differs from the bulk of scorers in the categorization of individuals. If we found such a case, we would tend to be suspicious of the quality/reliability of this scorer's work. As another example, we would like to detect sets of classification categories which encourage high levels of disagreement among scorers. For example, if two categories are ambiguously defined or confusing, scorers may be undecided about which category to place an individual.

The fact that a traditional measure of agreement, Cohen's kappa, will not detect the patterns of disagreement can be demonstrated with the following simplified example. Two scorers (S_1 and S_2) were required to categorize each of one hundred individuals into one of two categories (C_1 and C_2). Another two scorers were given the same task. The resul-s of the classifications are given below.

		Scorer 2		
		C1	C2	
Scorer 1	C1	25	25	50
	C2	25	25	50
		50	50	100

		Scorer 4		
		C1	C2	
Scorer 3	C1	0	0	0
	C2	50	50	100
		50	50	100

In both of the above tables the value of Cohen's kappa is zero. However, the two tables are clearly different. The first table could conceptually arise if each scorer randomly assigned people to one or the other class by flipping a coin. In the second table, scorer 3 never assigns anyone to the

first class. Because situations like the one exemplified in the second table are important to detect in practice, it is clear that some other procedure than kappa must be used to do so. Kappa, like other measures of agreement, largely ignores the information in the off-diagonal cells.

For the remainder of this paper we will turn our attention to the cases where the scorers disagree with each other while ignoring the cases where complete agreement occurs.

Measures of Diagreement

Statistics have been proposed to test patterns of scorer disagreement in several situations. The simplest case is two scorers and two categories. For example, suppose that the responses of N students to an item are being judged as either acceptable (A) or unacceptable (U) by two independent scorers. The response frequencies are displayed in the following 2 X 2 table.

		Scorer 2 A	Scorer 2 U	
Scorer 1	A	a	b	a+b
	U	c	d	c+d
		a+c	b+d	N = a+b+c+d

The null hypothesis that any disagreement bewteen scorers is due purely to random fluctuation can be tested with the chi-square test due to McNemar (1947) which, neglecting the correction for discontinuity, can be written $\chi^2 = \frac{(b-c)^2}{b+c}$.

Cochran (1950) generalized McNemar's test to the case of two categories and any number of scorers. Recently, Patil (1975) computed an exact distribution of Cochran's Q statistic.

Madansky (1963), in turn, generalized Cochran's test to the case of an arbitrary numer of categories. The procedures to be given in this paper extend Madansky's statistic in the manner suggested in Cochran's (1950) paper so that patterns of disagreement may be examined among subgroups of scorers and/or categories.

Before presenting the generalized tests, it should be noted that all of the statistics for examining patterns of disagreement given here are based on those cases where at least one scorer is in disagreement with the other scorers. Thus, the individuals for which all scorers were in complete agreement on categorization are ignored. In a simple example, with two scorers and two categories, the form of the test statistic is $Q = \frac{(b-c)^2}{b+c}$ using the notation of the previous table. Thus, the values of the main diagonal elements, a and d, where the scorers agreed do not enter in the computation.

Thus the tables

		Scorer 2 C1	Scorer 2 C2	
Scorer 1	C1	0	1325	1325
	C2	1175	0	1175
		1175	1325	2500

		Scorer 2 C1	Scorer 2 C2	
Scorer 1	C1	2000	5	2005
	C2	20	475	495
		2020	480	2500

each have identical values of Q = 9. However, in the first table, the two scorers never agreed with each other while in the second, the two scorers agreed ninety-nine percent of the time. Clearly, the implications of the two situations are different. We are apt to be more concerned about the disagreement in the first table than the disagreement in the second. For this reason, some examination of the amount of agreement, such as the proportion of cases where complete agreement occurs, needs to occur along with examination of the disagreement patterns.

The Test Statistics for Measuring Disagreement

Consider an experiment involving N individuals (subjects) and S scorers. Each scorer examines each individual and assigns that individual into one of C categories on the basis of some criteria. As an example, suppose that each individual responds to the same exercise. The response of an individual to that exercise is independently examined and assigned to a nominal response category by each scorer. Every individual is thus classified by every scorer on the basis of performance on the exercise. We wish to estimate the degree to which the categorization of an individual depends on the scorer doing the categorizing. We assume that the scoring of any individual is completely independent of the scoring of all other individuals. We want to examine the patterns of disagreement in the scoring of individuals. The null hypothesis for the tests to follow is that when disagreement between scorers occurs, this disagreement is purely random; that is, there is no pattern to the disagreement.

In the succeeding development, the following notation will be used. In all that follows, a "+" replacing an index represents summation over that index, a "." represents the mean over the replaced index.

(4) $d_{ijk} = \begin{cases} 1 & \text{if scorer } j \text{ assigns individual } i \text{ to category } k \\ 0 & \text{otherwise} \end{cases}$

(5) $d_{+jk} = \sum_{i=1}^{N} d_{ijk} =$ number of individuals assigned to category k by scorer j.

(6) $d_{i+k} =$ number of scorers who assigned individual i to category k.

(7) $d_{++k} =$ number of times any individual was assigned to category k by any (or all) of the scorers.

(8) $\bar{d}_{+\cdot k} = \frac{1}{S} \sum_{j=1}^{S} d_{+jk} =$ average number of individuals assigned to category k by a scorer.

A more precise statement of the null hypothesis to be tested is as follows:

> All classifications of an individual are entirely independent of any classification of any other individual. Further, for every category k given that d_{i+k} of the S scorers selected category k for individual i, every one of the scorers is assumed equally likely to have selected that category for that individual.

That is, the null hypothesis is that all scorers are interchangeable, that is the classification of an individual is independent of the scorer doing the classification. Let the vector $(k1,k2,\ldots,kS)_i$ represent the categories to which individual i is assigned by the scorers (so individual i is assigned to category k1 by scorer 1, category k2 by scorer 2, etc.). Then if the scorers are in fact interchangeable, all permutations of the above vector are equally likely to occur. For example, for S = 2 and C = 2, given that an individual is assigned once to category 1 and once to category 2, if the two scorers are interchangeable, the probability that he will be assigned to category 1 by scorer 1 is equal to the probability that scorer 2 assigns him to category 1.

As another example, suppose there are three categories and three scorers. Then the assignments (1,1,3), (1,3,1), (3,1,1) are all equally likely, as are the assignments (1,2,3), (1,3,2), (2,3,1), (2,1,3), (3,1,2) and (3,2,1). Note, however, that the assignments (1,1,3) and (1,2,3) do not have to be equally likely.

Based on this null hypothesis, Madansky (1963) demonstrated that the vector

$$(9) \quad (d_{+11},\ldots,d_{+S1},\ d_{+12},\ldots,d_{+S2},\ldots,d_{+1C},\ldots,d_{+SC}),$$

which has SC elements, asymptotically follows a multivariate normal distribution. The mean vector of this distribution is estimated by the vector

$(\bar{d}_{+\cdot 1},\ldots,\bar{d}_{+\cdot 1}, \bar{d}_{+\cdot 2},\ldots,\bar{d}_{+\cdot 2},\ldots,\bar{d}_{+\cdot C},\ldots,\bar{d}_{+\cdot C})$ and the covariance matrix is estimated by the SC X SC matrix Λ. A typical element of Λ is

$$\sigma_{ke} = \text{Cov}\ (d_{+jk},\ d_{+re})$$

which is constant for every value of j and r and which has the form

$$(10)\quad \sigma_{ke} = \begin{cases} \frac{1}{S}\, d_{++k} - \frac{1}{S^2}\, \Delta_{kk} & \text{if } k = e \\ -\frac{1}{S^2}\, \Delta_{ke} & \text{if } k \neq e \end{cases}$$

where

$$\Delta_{ke} = \sum_{i=1}^{N} d_{i+k}\, d_{i+e}.$$

Since $d_{+j+} = N$ for every j (because each scorer categorizes all N individuals) and since σ_{ke} is a constant for all scorers, it follows that the matrix Λ is singular of rank (S-1) (C-1).

Let

(11) Σ be the (C-1) x (C-1) matrix composed of the elements σ_{ke} given in (10) for k, e=1,...,C-1). The element in the ith row and jth column of Σ is σ_{ij}.

(12) V_{ke} = the element in the kth row and eth column of Σ^{-1}.

$$(13)\quad T_{ke} = \sum_{j=1}^{S} (d_{+jk} - \bar{d}_{+\cdot k})\ (d_{+j} - \bar{d}_{+\cdot e}).$$

Under the null hypothesis, the statistic

$$(14)\quad Q = \frac{S-1}{S} \sum_{k=1}^{C-1} \sum_{e=1}^{C-1} V_{ke}\, T_{ke}$$

is asymptotically distributed as χ^2 with (C-1) (S-1) degree of freedom.

Global Test

The hypothesis of interchangeability is tested with the statistic Q, given in equation (14). A significant value of Q indicates that not all scorers are interchangeable; that is, some scorer (or group of scorers)

tends to significantly differ in scoring patterns from the majority. Q as stated is a global test statistic which simultaneously compares the performance of all scorers among all classification categories. Subtests for comparing the performance between subgroups of scorers among all classification categories or for comparing the performance of some or all scorers within or between subsets of categories can be performed by recomputing Q for various subsets of the original data.

Comparisons Involving Subgroups of Scorers

If the global test statistic is significant, we may believe that this significance is largely due to one or more scorers performing differently from the rest. Suppose that the S scorers are partitioned into two groups. Let G(1) represent the first group of scorers while G(2) represents the second. Let S(1) be the number of scorers in G(1), and S(2) be the number of scorers in G(2), so that S(1) + S(2) = S. By the following partitioning of the global test statistic, we can test whether or not the two groups of scorers have different disagreement patterns. Further, the similarity of performance of all scorers within a group can be examined.

For group r = 1,2 let

(8a) $\bar{d}(r)_{+\cdot k} = \frac{1}{S(r)} \sum_{j \varepsilon G(r)} d_{+jk}$ = mean for category k over scorers in group G(r).

(13a) $T(r)_{ke} = \sum_{j \varepsilon G(r)} (d_{+jk} - \bar{d}(r)_{+\cdot k})(d_{+je} - \bar{d}(r)_{+\cdot e})$ = corrected sum of products of scorer totals within group G(r) between categories k and e.

(13b) $T(1,2)_{ke} = \frac{S(1)\,S(2)}{S} (\bar{d}(1)_{+\cdot k} - \bar{d}(2)_{+\cdot k})(\bar{d}(1)_{+\cdot e} - \bar{d}(2)_{+\cdot e})$ = corrected sum of products between groups between categories k and e.

(Note that $T_{ke} = T(1)_{ke} + T(2)_{ke} + T(1,2)_{ke}$.) To test for differences in patterns of scoring between group 1 as a whole and group 2 as a whole, compute

$$(14a) \quad Q(1,2) = \frac{(S-1)}{S} \sum_{k=1}^{C-1} \sum_{e=1}^{C-1} V_{ke} \cdot T(1,2)_{ke}$$

and compare it with χ^2 with (C-1) degrees of freedom. To test for differences in patterns of scoring within group G(r) alone, given that Q(1,2) was not significant, compute

$$(14b) \quad Q(r) = \frac{(S-1)}{S} \sum_{k=1}^{C-1} \sum_{e=1}^{C-1} V_{ke} \cdot T(r)_{ke}$$

and compare with χ^2 with (C-1) (S(r)-1) degrees of freedom.

The matrix Σ of (11) was constructed as a measure of internal variability under the assumption that the conditional probability that a specific individual is assigned to a given category by a particular scorer, given he was assigned to that category by some scorer, is the same for <u>all</u> scorers. Given that there is no difference in scoring patterns between G(1) and G(2), Σ is valid to use in the construction of Q(r). However, if Q(1,2) is significant, Σ cannot be considered a valid estimate since the conditional probability of assignment of a given individual to a given category is no longer constant across scorers. In this case, the variability must be estimated for each group of scorers separately.

Let $\Sigma(r)$ be the estimate computed only on data relating to scorers in group G(r). Except for restricting the data, $\Sigma(r)$ is calculated in the same manner as Σ. Let $V(r)_{ke}$ be the k^{th} element of $(\Sigma(r))^{-1}$. If Q(1,2) is found to be significant, compute Q(r) as

$$(14b^*) \quad Q^*(r) = \frac{(S(r)-1)}{S(r)} \sum_{k=1}^{C-1} \sum_{e=1}^{C-1} V(r)_{ke} \cdot T(r)_{ke}$$

which is compared with χ^2 with (C-1) (S(r)-1) degrees of freedom.

The above can easily be extended to cases with more than two groups of scorers.

Comparisons Involving Aggregates of Categories

Suppose the categories are partitioned into q category classes. Denote the kth such class by K(k), let the number of categories in this class be C(k). For example, response categories could be partitioned into two classes, acceptable responses in class K(1) and unacceptable responses in class K(2). Two scorers will be said to agree in the categorization of an individual in terms of the category classes if they each assign the individual to categories which belong to the same class.

To compare the disagreement patterns of scorers in terms of category classes, collapse the original data into these classes and proceed in the manner previously given. All tests and procedures previously given are unchanged except that any reference to a category is now a reference to a category class. Any reference to C, the number of original categories, is now a reference to q, the number of category classes. Any reference to d_{ijk} is now a reference to Σd_{ije} where the summation is taken over all exercises e which are in the class K(k).

To compare disagreement within a particular category class, delete from the original data any individual who was not assigned to that class by all scorers. On this reduced set of data, proceed as before replacing any reference to C by C(k), the number of categories in the class in question.

Applying the Tests to Data from Multi-Way Tables

Although the various statistics have been presented in a form which utilized the distribution of classifications by scorers on a person-by-person basis, it should be noted that this level of detail in the data is not necessary to perform the tests. All the given tests can be conducted on a data set which is an S-way table of frequencies. In this S-way table, each dimension (rows, columns, levels, etc.) represents a scorer while the categories within each dimension are the C classification categories to be used by the scorers. The basic data are S-way cell counts of the form $n_{(k1,k2,\ldots,kS)}$ where $n_{(k1,k2,\ldots,kS)}$ is the total number of individuals which were assigned to category k_1 by scorer 1, to category k_2 by scorer 2,..., and to category k_S by scorer S.

The relationships between these S-way cell counts and the aggregates needed to perform the tests are as follows:

d_{+jk} = the sum of all cell frequencies $n_{(k1,\ldots,ks)}$ whose indices have the value k in the jth position. For example, with S=4,

$$d_{+2k} = n_{(+,k,+,+)} = \sum_{i_1=1}^{C} \sum_{i_3=1}^{C} \sum_{i_4=1}^{C} n_{(i_1,k,i_3,i_4)}$$

$$d_{++k} = n_{(k,+,\ldots,+)} + n_{(+,k,+,\ldots,+)} + \ldots + n_{(+,\ldots,+,k)}$$

$$\Delta_{ke} = \begin{cases} \sum_{h=1}^{S-1} \sum_{i=1}^{h} (S-h)(i)\, g_{ke}(S-h,i) & \text{for } k \neq e \\ \sum_{h=1}^{S} (h)^2\, g_{kk}(h,h) & \text{for } k=e \end{cases}$$

where $g_{ke}(h,i)$ is the number of individuals who were assigned to category k by exactly h scorers and to category e by exactly i scorers and is the summation of all cell frequences which have exactly h indices equal to k and exactly i indices equal to e. For example, with S=3 and C=3,

$$g_{ke}(1,2) = n(k,e,e) + n(e,k,e) + n(e,e,k)$$
$$g_{11}(2,3) = n(1,1,2) + n(1,2,1) + n(2,1,1) + n(1,1,3) + n(1,3,1) + n(3,1,1)$$

Example

S=3 scorers, C=3 categories, N=165 individuals

Table 1. Values of n(i,j,k)

	Scorer 3 Picks Category 1				Scorer 3 Picks Category 2				Scorer 3 Picks Category 3			
	Category				Category				Category			
	Scorer 2				Scorer 2				Scorer 2			
Scorer 1	1	2	3	n(i,+,1)	1	2	3	n(i,+,2)	1	2	3	n(i,+,3)
1	12	3	9	24	0	4	5	9	8	2	4	14
2	3	6	5	14	4	27	2	33	2	0	3	5
3	2	7	9	18	5	3	0	8	3	4	33	40
n(+,j,k)	17	16	23	56	9	34	7	50	13	6	40	59

Example: n(3,1,1) = 2; n(2,2,2) = 27
n(1,+,1) = 24; n(+,+,1) = 56
n(1,+,+) = 24 + 9 + 14 = 47
n(+,1,+) = 17 + 9 + 13 = 39

Table 2. Values of d_{+jk}

	Category (k)		
Scorer (j)	1	2	3
1	47	52	66
2	39	56	70
3	56	50	59
Total over Scorers (d_{++k})	142	158	195
Mean over Scorers ($\bar{d}_{+\cdot k}$)	47.33	52.67	65

Table 3. Values of $g_{ke}(h,i)$

Category k	Category e	Number of Scorers h: 1 i: 1	h: 1 i: 2	h: 2 i: 1	h: 2 i: 2	h: 3 i: 3
1	1	56	--	--	25	12
1	2	26	14	6	--	--
1	3	26	16	19	--	--
2	1	26	6	14	--	--
2	2	39	--	--	19	27
2	3	26	7	5	--	--
3	1	26	19	16	--	--
3	2	26	5	7	--	--
3	3	50	--	--	23	33

Example: $g_{13}(1,2) = g_{31}(2,1) = n(1,3,3) + n(3,1,3) + n(3,3,1) = 4 + 3 + 9 = 16$

Table 4. Values of T_{ke}, Δ_{ke}, σ_{ke}

Statistic	Category* k: 1 e: 1	k: 1 e: 2	k: 1 e: 3	k: 2 e: 2	k: 2 e: 3	k: 3 e: 3
T_{ke}	144.67	-50.67	-94	18.67	32	62
Δ_{ke}	264	66	96	358	50	439
σ_{ke}	18.000	- 7.333	-10.667	12.889	- 5.556	16.222

*All statistics are symmetric so that, for example, $T_{ke} = T_{ek}$.

Example: $\Delta_{11} = (1)^2.56 + (2)^2.25 + (3)^2.12 = 264$
$\Delta_{12} = (2)(1).6 + (1)(1).26 + (1)(2).14 = 66$

Global Test

Using categories 1 and 2 (the results are identical using any pair of categories).

$$\Sigma = \begin{bmatrix} 18.000 & - 7.333 \\ - 7.333 & 12.889 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} = \begin{bmatrix} .0723 & .0411 \\ .0411 & .1010 \end{bmatrix}$$

$$Q = 2/3 \{V_{11}T_{11} + V_{12}T_{12} + V_{21}T_{21} + V_{22}T_{22}\} = 5.45$$
with $(S-1)(C-1) = (2)(2) = 4$ degrees of freedom.

Testing Subgroups of Scorers

Using all classification categories, we will compare scorer 2 with the combined performance of scorers 1 and 3. Thus

$G(1)$ = {scorer 2} $\quad G(2)$ = {scorers 1 and 3}
$S(1) = 1 \quad S(2) = 2$

$\bar{d}(1)_{+\cdot 1} = d_{+21} = 39 \quad \bar{d}(1)_{+\cdot 2} = 56$
$\bar{d}(2)_{+\cdot 1} = (d_{+11} + d_{+21})/2 = 51.5$
$\bar{d}(2)_{+\cdot 2} = 51$

Table 5. Values of $T(1)_{ke}$, $T(2)_{ke}$, $T(1,2)_{ke}$

	Categories		
k:	1	1	2
Statistic e:	1	2	2
$T(1)_{ke}$	0	0	0
$T(2)_{ke}$	40.5	- 9	2
$T(1,2)_{ke}$	104.17	-41.67	16.67

To compare scorer 2 with scorers 1 and 3

$$Q(1,2) = 2/3 \{V_{11} T(1,2)_{11} + 2 V_{12} T(1,2)_{12} + V_{22} T(1,2)_{22}\}$$

$= 3.86$ with $(C-1) = 2$ degrees of freedom.

Since $Q(1,2)$ is nonsignificant, we can assume that Σ is a valid measure of the internal variability of group $G(2)$. Thus to compare scorer 1 with scorer 3, we compute

$$Q(2) = 2/3 \sum_{k=1}^{2} \sum_{e=1}^{2} V_{ke} T(2)_{ke} = 1.59 \quad 2 \text{ degrees of freedom}$$

To demonstrate the technique, we recompute the statistic $Q(2)$ as if $Q(1,2)$ were found to be significantly large. In this case, to compare scorer 1 with scorer 3, the variance matrix Σ must be recomputed using only data from scorers 1 and 3. This can be conveniently obtained from the following table.

Table 6. Values of $n(i,+,k)$

Scorer 1	Scorer 3 Category 1	2	3	$n(i++)$
1	24	9	14	47
2	14	33	5	52
3	18	8	40	66
$n(++k)$	56	50	59	165

The following can be obtained from this table in a manner analagous to previous computations.

$d(2)_{++1} = 103$ $\qquad$ $d(2)_{++2} = 102$

$\Delta(2)_{11} = 151 = 2^2(24) + 1^2(9 + 14 + 14 + 18)$

$\Delta(2)_{12} = 23 = 1 \cdot 1\ (14 + 9)$

$\Delta(2)_{22} = 168$

$\sigma(2)_{11} = 1/2(103) - 1/4(151) = 13.75$

$\sigma(2)_{12} = -5.75$

$\sigma(2)_{22} = 9.00$

$$\Sigma(2)^{-1} = \begin{vmatrix} .0992 & .0634 \\ .0634 & .1516 \end{vmatrix} = \begin{vmatrix} V(2)_{11} & V(2)_{12} \\ V(2)_{21} & V(2)_{22} \end{vmatrix}$$

The test statistic for comparing scorer 1 with scoer 3 is

$$Q^*(2) = \frac{(2-1)}{2}(3.18) = 1.59 \text{ with } (3-1)(2-1) = 2 \text{ degrees of freedom}$$

Tests Involving Aggregates of Categories

We compare the performance of scorers between the two category classes: $K(1)$ = {category 1} and $K(2)$ = {categories 2 and 3}.

The equivalent of Table 1 for these classes is

Scorer 3 Picks Class 1

Scorer 1	Scorer 2 Class 1	Class 2
Class 1	12	12
Class 2	5	27

Scorer 3 Picks Class 2

	Scorer 2 Class 1	Class 2
	8	15
	14	72

The analogue of tables 2 and 4 is

Category Class (k)	d_{+1k}	d_{+2k}	d_{+3k}	d_{++k}	T_{kk}	Δ_{kk}	σ_{kk}
1	47	39	56	142	144.67	264	18
2	118	126	109	353	--	--	--

The test statistic for comparing category classes 1 and 2 is

$$Q = 2/3 \frac{T_{11}}{\sigma_{11}} = 5.36 \text{ with } (2-1)(3-1) = 2 \text{ degrees of freedom.}$$

REFERENCES

Burton, N. "The Effect of Position and Format on the Difficulty of Assessment Exercises." Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Burton, J. and Johnson, E. "Generalizability of Nominal Categories." In preparation, 1978.

Cochran, W. G. "The Comparision of Percentages in Matched Samples," Biometrika, 37 (1950), 256-266.

Coffman, W. E. "Essay Examinations," Educational Measurement, second edition. Edited by R. L. Thorndike. Washington, D.C.: American Council on Education, 1971.

Cohen, J. A. "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, 20 (1960), 37-46

Cronbach, L., Gleser, G., and Rajaratnum, N. "The Dependability of Behavioral Measurements." New York: Wiley, 1972.

D'Agostino, R. B. "A Second Look at Analysis of Variance on Dichotomous Data," Journal of Educational Measurement, 8 (1971), 327-333.

Davidson, T. N. "Coding Reliability and Consensus Coding." Unpublished manuscript, Survey Research Center, University of Michigan, circa 1973.

Folsum, R. E. "National Assessment Approach to Sampling Error Estimation," Sampling Error Monograph 25U-796-5, Research Triangle Institute, Research Triangle Park, North Carolina, April 1977.

Glass, G. V and Stanley, J. Statistical Method in Education and Psychology. Englewood Cliffs, New Jersey: Prentice-Hall, 1970.

Goodman, L. A. and Kruskal, W. H. "Measures of Association for Cross-Classifications," Journal of the American Statistical Association, 49 (1954), 732-764.

Kendall, M. G and Stuart, A. The Advanced Theory of Statistics. Second edition. London: Charles Griffin and Co., 1967.

Kruppindorff, K. "Bivariate Agreement Coefficients for Reliability of Data," Sociological Methodology. Edited by E. F. Borgatta and G. W. Bohrnstedt. San Francisco: Jossey-Bass, 1970.

Light, R. J. "Issues in the Analysis of Qualitative Data," Second Handbook of Research on Teaching. Edited by RMW Travers. Chicago: Rand McNally, 1973.

McNemar, W. "Note on the Sampling Error of Difference Between Correlated Proportions or Percentages," Psychometrika, 12, (1947), 153-157.

Madansky, A. "Tests of Homogeneity for Correlated Samples," Journal of the American Statistical Association, 58 (1963), 97-119.

Patil. K. D. "Cochran's Q Test: Exact Distribution," Journal of the American Statistical Association, 70 (1975).

Pearson, K. "Mathematical Contribution to the Theory of Evaluation," IX, Part I. Philosophical Transactions of the Royal Society. London: Series A, 197 (1901), 285-479.

Scott, W. A. "Reliability of a Content Analysis: The Case of Nominal Scale Coding," Public Opinion Quarterly, 19 (1955), 321-325.